

# A data-driven basis for direct estimation of functionals of distributions

Alan Wisler, Visar Berisha, Andreas Spanias, Alfred O. Hero

**Abstract**—A number of fundamental quantities in statistical signal processing and information theory can be expressed as integral functions of two probability density functions. Such quantities are called density functionals as they map density functions onto the real line. For example, information divergence functions measure the dissimilarity between two probability density functions and are particularly useful in a number of applications. Typically, estimating these quantities requires complete knowledge of the underlying distribution followed by multi-dimensional integration. Existing methods make parametric assumptions about the data distribution or use non-parametric density estimation followed by high-dimensional integration. In this paper, we propose a new alternative. We introduce the concept of “data-driven” basis functions - functions of distributions whose value we can estimate given only samples from the underlying distributions without requiring distribution fitting or direct integration. We derive a new data-driven complete basis that is similar to the deterministic Bernstein polynomial basis and develop two methods for performing basis expansions of functionals of two distributions. We also show that the new basis set allows us to approximate functions of distributions as closely as desired. Finally, we evaluate the methodology by developing data driven estimators for the Kullback-Leibler divergences and the Hellinger distance and by constructing tight data-driven bounds on the Bayes Error Rate.

**Index Terms**—Divergence estimation, direct estimation, nearest neighbor graphs, Bernstein polynomial

## I. INTRODUCTION

Information divergence measures play a central role in the fields of machine learning and information theory. Information divergence functions, functionals that map density functions to  $\mathbb{R}$ , have been used in many signal processing applications involving classification [1], segmentation [2], source separation [3], clustering [4], and other domains. In machine learning, a sub-class of these divergences known as  $f$ -divergences [5], are widely used as surrogate loss functions since they form convex upper bounds on the non-convex 0-1 loss [6].

Although these measures prove useful in a variety of applications, the task of estimating them from multivariate probability distributions using finite sample data can pose a significant challenge. The literature shows that there are three general classes of methods for estimating divergence [7]: 1) parametric methods, 2) non-parametric methods based on density estimation, and 3) non-parametric methods based on direct (or graph-based) estimation. Parametric methods are the most common choice for estimation, and typically offer good convergence rates ( $1/N$ ) when an accurate parametric model is selected. The fundamental limitation of parametric methods, is that an accurate parametric model is rarely available in real world problems and using an inaccurate parametric model can heavily bias the final estimate. As an alternative, when no parametric form is known, non-parametric density estimates such as kernel density estimation [8], histogram estimation [9], or  $k$ -nearest neighbor density estimation [10], allow density estimates to be formed without assuming a parametric form for the underlying distribution. While these methods are quite powerful in certain scenarios, they are generally high variance, sensitive to outliers, and scale poorly with dimension [7]. An alternative to these two classes of methods, is direct (or

graph-based) estimation, which exploits the asymptotic properties of minimal graphs in order to *directly* estimate distribution functionals without ever estimating the underlying distributions themselves. These methods have been used to estimate density functionals such as entropy [11], the  $\alpha$ -divergence [7], and the  $D_p$ -divergence [12]. This class of methods can have faster asymptotic convergence rates [7] and bypass the complication of fine tuning parameters such as kernel width or histogram bin size.

A fundamental limitation of previously-derived direct estimation methods is the specificity of the estimation method to the density functional being estimated. With plug-in estimators, the same general approach can be used to estimate any function of distributions, whereas the graph-based estimation methods have been limited to a specific divergence or class of divergences. In constructing direct estimators, authors typically analyze an asymptotic property of a graph-theoretic quantity and scale it appropriately to generate an estimator for a given information-theoretic measure. The resulting estimator is only applicable to a particular density functional and cannot be generalized to other functionals. Our previous work is an example of this [12]. We previously introduced a new divergence measure that can be directly estimated from the Friedman-Rafsky statistic [13], however there is no direct way to generalize the same approach to estimate other density functionals. In this paper, we attempt to resolve this issue by introducing a general approach for estimating a wide form of distribution functionals. To accomplish this, we propose a complete set of “data-driven” basis functions. We term these basis functions “data-driven” if we can estimate their value given only samples from the underlying distributions without distribution fitting or direct integration. We show that a rather broad class of distribution functionals, which includes the family of  $f$ -divergences, can be approximated as closely as desired through linear combinations of our proposed basis.

In the next section, we review the literature in this area. In Section III we provide a detailed description of the problem this paper attempts to solve and establish some of the basic mathematical notation used throughout this paper. In Section IV we introduce a set of graph-theoretic basis functions as well as prove that a wide range of information theoretic quantities can be represented by a linear combination of functions in this set. In sections V and VI-B, we explore the limitations of the proposed methodology in the finite sample regime and propose two alternate fitting routines to identify weights to map these basis functions to quantities of interest. In section VII we empirically investigate how the proposed method can be used to estimate popular divergence measures (the KL-divergence, the Hellinger distance, the  $D_p$ -divergence), and we compare its performance to various parametric and non-parametric alternatives. In Section VIII, we show how the method can be extended to form tighter bounds on the Bayes error rate for binary classification problems. Section IX offers some concluding remarks.

## II. RELATED WORK

In the majority of cases scientists use closed form solutions for different distribution types when approaching problems in statistical signal processing and communications [14]. However,

in the statistical learning literature a number of non-parametric estimators for entropy and divergence functions have been proposed. Graph-based estimators of the Rényi entropy were introduced by Hero and Costa in [15]. Other studies have built upon this work to develop non-parametric estimators for both the Shannon and Rényi entropy [16]–[20].

A general method for non-parametric estimation involves histogram binning followed by plug-in estimation [21], [22]. When the bin-size is adjusted as a function of the number of available samples per bin, this histogram plug-in method is known as Grenander's method of sieves and it enjoys attractive non-parametric convergence rates [23], [24]. While these methods may work well for small data dimension ( $d = 1, 2$ ), their complexity becomes prohibitive for larger dimensions. A related approach involves non-parametric density estimation followed by a plug-in estimate of the quantity of interest [25]–[27]. Sricharan and Hero investigated the bias and variance associated with plug-in estimates of functions of densities [25], [26]. Density estimators usually have tunable parameters that must be correctly set using, for example, cross-validation. In contrast to these methods, the approaches we propose here do not require density estimation.

More recently, estimates of divergence functions that rely on estimates of the likelihood ratio instead of density estimation have been proposed for estimating the  $\alpha$ -divergence and the  $L_2$ -divergence [6], [28]–[31]. These methods estimate the likelihood ratio of the two density functions and plug that value into the divergence functions. Our method aims to go beyond estimating only divergence measures. We propose an extensible framework for constructing estimators for different quantities by using data-driven basis functions that avoid plug-in approaches involving density estimation.

Bounds on optimal performance are a key component in the statistical signal processing literature. For classification problems, it is often desirable to bound the Bayes error rate (BER) - the minimum achievable error in classification problems. The well-known Chernoff upper bound on the probability of error has been used in a number of statistical signal processing applications [32]. It motivated the Chernoff  $\alpha$ -divergence [7]. The Bhattacharyya distance, a special case of the Chernoff  $\alpha$ -divergence for  $\alpha = \frac{1}{2}$ , upper and lower bounds the BER [33], [34]. Beyond the bounds on the BER based on divergence measures, a number of other bounds exist based on other functionals of the distributions [35], [36]. For estimation problems, the Fisher information matrix (FIM) bounds the variance of the optimal unbiased estimator (through its relationship with the CRLB). The authors have also previously introduced the  $D_p$  divergence, a non-parametric  $f$ -divergence, and showed that it provides provably tighter bounds on the BER than the BC bound [12]. They extended this work to estimation of the Fisher information in [37].

Our data-driven basis, consisting of Bernstein polynomials, can be used to estimate functionals of distributions and to bound optimal performance in classification. Bernstein polynomials of a different form have been used for density estimation [38]–[43]. In contrast to this work, our methods do not rely on density estimation. In fact, the data-driven basis functions we propose bypass density estimation altogether and directly estimate the value of a parameter of interest.

### III. PROBLEM SETUP

In this section, we will set up the problem and establish the notation that will be used throughout the rest of the paper. We are given a set of data  $[\mathbf{X}, \mathbf{y}]$  containing  $N$  instances, where each instance is represented by a  $d$ -dimensional feature vector  $\mathbf{x}_i$  and a binary label  $y_i$ . Suppose that this data is sampled from underlying distribution,  $f_{\mathbf{x}}(\mathbf{x})$ , where

$$f_{\mathbf{x}}(\mathbf{x}) = p_0 f_0(\mathbf{x}) + p_1 f_1(\mathbf{x}) \quad (1)$$

is made up of the two conditional class distributions  $f_0(\mathbf{x})$  and  $f_1(\mathbf{x})$  for classes 0 and 1, with prior probabilities  $p_0$  and  $p_1$  respectively. As a simple application of Bayes theorem, we can define the posterior likelihood of class 1,  $\eta(\mathbf{x})$ , evaluated at a point  $\mathbf{x} = \mathbf{x}^*$ , as

$$\begin{aligned} \eta(\mathbf{x}^*) &= P[y = 1 | \mathbf{x} = \mathbf{x}^*] = \frac{P[y = 1] f_{\mathbf{x}}(\mathbf{x}^* | y = 1)}{f_{\mathbf{x}}(\mathbf{x}^*)} \\ &= \frac{p_1 f_1(\mathbf{x}^*)}{f_{\mathbf{x}}(\mathbf{x}^*)} = \frac{p_1 f_1(\mathbf{x}^*)}{p_0 f_0(\mathbf{x}^*) + p_1 f_1(\mathbf{x}^*)} \end{aligned} \quad (2)$$

We can similarly define the posterior probability for class 0 as

$$P[y = 0 | \mathbf{x} = \mathbf{x}^*] = \frac{p_0 f_0(\mathbf{x}^*)}{p_0 f_0(\mathbf{x}^*) + p_1 f_1(\mathbf{x}^*)}, \quad (3)$$

and since  $y$  is binary,

$$P[y = 0 | \mathbf{x} = \mathbf{x}^*] = 1 - \eta(\mathbf{x}^*). \quad (4)$$

To simplify the notation, we remove the dependence of  $\eta$  on  $\mathbf{x}^*$  in the analysis that follows.

Suppose that we wish to estimate some functional  $G(f_0, f_1)$  of distributions  $f_0(\mathbf{x})$  and  $f_1(\mathbf{x})$ , which can be expressed in the following form

$$G(f_0, f_1) = \int g(\eta) f_{\mathbf{x}}(\mathbf{x}) d\mathbf{x}. \quad (5)$$

Throughout the rest of this paper, we will refer to the form presented in (5) as the *standard* form. Many functionals in machine learning and information theory, such as  $f$ -divergences and loss functions, can be expressed this way. Consider the family of  $f$ -divergences as an example. They are defined as

$$D_{\phi}(f_0, f_1) = \int \phi\left(\frac{f_0(\mathbf{x})}{f_1(\mathbf{x})}\right) f_1(\mathbf{x}) d\mathbf{x}, \quad (6)$$

where  $\phi(t)$  is a convex or concave function unique to the given  $f$ -divergence. By substituting

$$\frac{f_0(\mathbf{x})}{f_1(\mathbf{x})} = \frac{p_1(1 - \eta)}{p_0 \eta} \quad (7)$$

and

$$f_1(\mathbf{x}) = \frac{\eta}{p_1} f_{\mathbf{x}}(\mathbf{x}) \quad (8)$$

we can redefine (6) as

$$D_{\phi}(f_0, f_1) = \int \phi\left(\frac{p_1(1 - \eta)}{p_0 \eta}\right) \frac{\eta}{p_1} f_{\mathbf{x}}(\mathbf{x}) d\mathbf{x}. \quad (9)$$

Thus any  $f$ -divergence can be presented in the standard form simply by defining the generator function  $g(\eta)$  as

$$g(\eta) = \frac{\eta}{p_1} \phi\left(\frac{p_1(1 - \eta)}{p_0 \eta}\right). \quad (10)$$

While estimating these types of divergence functionals has traditionally relied on plug-in methods [44], we propose an alternative procedure which bypasses density estimation. We do this by representing the functional in terms of the asymptotic limit of a linear combination of graph-theoretic basis functions.

Suppose that there exist a set of basis functions  $H_1(\eta), \dots, H_K(\eta)$  that can be similarly expressed as

$$H_i(f_0, f_1) = \int h_i(\eta) f_{\mathbf{x}}(\mathbf{x}) d\mathbf{x}. \quad (11)$$

If we assume that there exist a set of coefficients such that

$$g(\eta) \approx \sum_{i=1}^K w_i h_i(\eta), \quad (12)$$

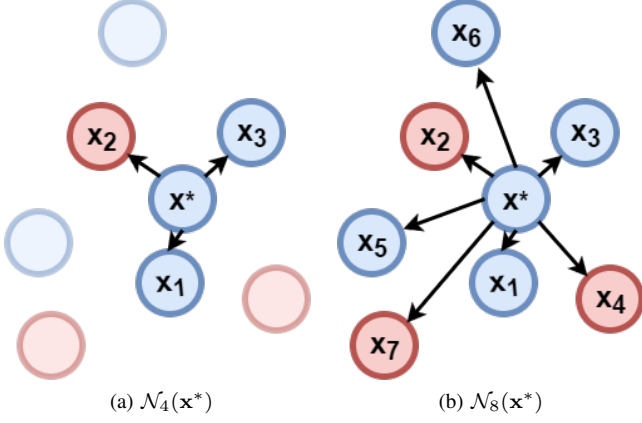


Fig. 1: Illustration of two neighborhoods of  $\mathbf{x}^*$  for  $k=4$  and  $k=8$ , instances with  $y=0$  are blue while instances with  $y=1$  are red. In the first scenario  $\Phi_4(\mathbf{x}^*)=1$ , since only one instance in  $\mathcal{N}_4(\mathbf{x}^*)$  is red. In the second scenario  $\Phi_8(\mathbf{x}^*)=3$ , since three of the eight instance in  $\mathcal{N}_8(\mathbf{x}^*)$  are red.

then consequently

$$G(f_0, f_1) \approx \hat{G}(f_0, f_1) = \sum_{i=1}^K w_i H_i(\eta), \quad (13)$$

where the sense of approximation is that the  $\ell_2$  norm of the difference between the right and left hand sides is small. In the following section, we will introduce a set of basis functions that have the desired properties.

#### IV. GRAPH-THEORETIC BASIS FUNCTIONS

Consider the dataset  $[\mathbf{X}, \mathbf{y}]$  previously defined. Suppose we select an arbitrary instance  $\mathbf{x}^*$  from  $\mathbf{X}$  and examine it along with the set of its  $k-1$  nearest neighbors  $\mathbf{x}_{NN}^1, \mathbf{x}_{NN}^2, \dots, \mathbf{x}_{NN}^{k-1}$  in  $\mathbf{X}$ . We can define the neighborhood set  $\mathcal{N}_k(\mathbf{x}^*) = [\mathbf{x}^*, \mathbf{x}_{NN}^1, \dots, \mathbf{x}_{NN}^{k-1}]$ , as the union of  $\mathbf{x}^*$  and its  $k-1$  nearest neighbors. Using this we define  $\Phi_k(\mathbf{x}^*)$  as the number of instances in the neighborhood set which are drawn from class 1, or alternatively, the sum of  $y$  across all points in the neighborhood

$$\Phi_k(\mathbf{x}^*) = \sum_{i: \mathbf{x}_i \in \mathcal{N}_k(\mathbf{x}^*)} y_i. \quad (14)$$

Figure 1 provides a simple illustration to help explain how  $\Phi_k(\mathbf{x}^*)$  is calculated. Calculating  $\Phi_k$  is similar to how nearest neighbor classifiers make decisions, but with two important differences:

- 1) The base instance  $\mathbf{x}^*$  is considered in the neighborhood indistinguishably from other instances in  $\mathcal{N}_k(\mathbf{x}^*)$
- 2) Where traditional  $k$ -NN classifiers are concerned only with identifying the majority, we are interested in the exact number of instances drawn from each class

In essence  $\Phi_k(\mathbf{x}^*)$  tells us something about the probability that  $y=1$  for instances on or near  $\mathbf{x}^*$ . Since we are more concerned with the dataset as a whole than the local characteristics in  $\mathbf{x}$ , we define the statistic  $\rho_{r,k}$  to be the fraction of instances  $\mathbf{x} \in \mathbf{X}$ , for which  $\Phi_k(\mathbf{x}) = r$ ,  $r \leq k$ . If we define the indicator function  $\theta_{r,k}(\mathbf{x})$  as

$$\theta_{r,k}(\mathbf{x}) = \begin{cases} 1 & \Phi_k(\mathbf{x}) = r \\ 0 & \text{otherwise,} \end{cases} \quad (15)$$

then this test statistic  $\rho_{r,k}$  can be represented by

$$\rho_{r,k}(\mathbf{X}) = \frac{1}{N} \sum_{\mathbf{x} \in \mathbf{X}} \theta_{r,k}(\mathbf{x}). \quad (16)$$

The function  $\rho_{r,k}(\mathbf{X})$  is simply the proportion of  $k$ -NN neighborhoods that contain exactly  $r$  points from class  $y=1$ . This statistic has a number of desirable qualities. We show that this statistic asymptotically converges to a function of the underlying distributions that can be described in the form outlined in (5). The following is proven in Appendix A.

**Theorem 1.** For any finite  $k$ , as the number of samples ( $N$ ) approaches infinity,

$$\lim_{N \rightarrow \infty} \rho_{r,k} = \int \binom{k}{r} \eta^r (1-\eta)^{k-r} f_{\mathbf{x}}(\mathbf{x}) d\mathbf{x},$$

almost surely.

We propose to use the asymptotic form of  $\rho_{r,k}$  defined in Theorem 1 as a basis function for estimating functionals of the form (5),

$$H_{r,k}(f_0, f_1) = \lim_{N \rightarrow \infty} \rho_{r,k} \quad (17)$$

$$= \int h_{r,k}(\eta) f_{\mathbf{x}}(\mathbf{x}) d\mathbf{x} \quad (18)$$

where

$$h_{r,k}(\eta) = \binom{k}{r} \eta^r (1-\eta)^{k-r}. \quad (19)$$

The function (19) is the  $r^{\text{th}}$  Bernstein basis polynomial of degree  $k$  [45]. Bernstein's proof of the Weierstrass Approximation Theorem [46] asserts that any continuous function  $g(\eta)$  can be uniformly approximated on  $\eta \in [0, 1]$  to any desired accuracy by a linear combination functions in (19) of the form

$$g(\eta) = \sum_{r=0}^k g\left(\frac{r}{k}\right) h_{r,k}(\eta), \quad (20)$$

for  $k$  sufficiently large.

Combining this result with Theorem 1, we can show that as  $N \rightarrow \infty$  and  $k \rightarrow \infty$  in a linked manner such that  $\frac{k}{N} \rightarrow 0$ , any function that can be represented in the form (5) can be expressed as a linear combination of  $\rho_{r,k}$ .

**Theorem 2.** As  $N \rightarrow \infty$  and  $k \rightarrow \infty$  in a linked manner such that  $\frac{k}{N} \rightarrow 0$ , any function  $G(f_0, f_1)$  that can be expressed in the form

$$G(\eta) = \int g(\eta) f_{\mathbf{x}}(\mathbf{x}) d\mathbf{x}$$

can be represented by

$$G(f_0, f_1) = \sum_{r=0}^k g\left(\frac{r}{k}\right) \rho_{r,k} \quad (21)$$

almost surely.

Theorem 2 provides an asymptotically consistent method of estimating a variety of information-theoretic functions that makes no assumptions on the underlying distributions and can be calculated without having to perform density estimation. Throughout the rest of the paper we will refer to the weights  $g(k/r)$  in the approximation specified by (21) in Theorem 2 as the Bernstein weights. We next turn to the finite sample properties of the estimator  $\rho_{k,r}$ .

#### V. FINITE SAMPLE CONSIDERATIONS

The previous Section investigated the asymptotic properties of linear combinations of the proposed set of empirically estimable

basis functions. The asymptotic consistency of the proposed method is valuable, however in real world scenarios, data is inherently a finite resource, and as a result the efficacy of this method is heavily dependent on its convergence characteristics in the finite sample regime. In this section, we will take a detailed look into how restricting both  $N$  and  $k$  affects our ability to estimate functions of two distributions. To do this, it is necessary to first break down the different sources of error in the proposed methodology.

#### A. Estimation vs. Approximation Error

The goal of this paper is to empirically estimate the functional  $G(f_0, f_1)$  of the two underlying distributions  $f_0(\mathbf{x})$  and  $f_1(\mathbf{x})$  using a linear combination of directly estimable basis functions

$$\hat{G}(f_0, f_1) = \sum_{r=0}^k w_r \hat{H}_{r,k}(f_0, f_1). \quad (22)$$

We can define the error in this estimate as

$$\begin{aligned} e_T &= G(f_0, f_1) - \hat{G}(f_0, f_1) \\ &= G(f_0, f_1) - \sum_{r=1}^k w_r \hat{H}_{r,k}(f_0, f_1). \end{aligned} \quad (23)$$

From Theorem 2, we know that if the weights  $w_r$  are defined as the Bernstein weights  $g(r/k)$  then

$$\lim_{\substack{N \rightarrow \infty, k \rightarrow \infty \\ k/N \rightarrow 0}} e_T = 0. \quad (24)$$

Suppose that we keep  $k$  fixed while  $N \rightarrow \infty$ . In this scenario

$$\begin{aligned} \lim_{N \rightarrow \infty} e_T &= G(f_0, f_1) - \lim_{N \rightarrow \infty} \sum_{r=1}^k w_r \hat{H}_{r,k}(f_0, f_1) \\ &= G(f_0, f_1) - \sum_{r=1}^k w_r H_{r,k}(f_0, f_1) \\ &= \int g(\eta) f_{\mathbf{x}}(\mathbf{x}) d\mathbf{x} - \sum_{r=0}^k w_r \int h_{r,k}(\eta) f_{\mathbf{x}}(\mathbf{x}) d\mathbf{x} \\ &= \int \left[ g(\eta) - \sum_{r=0}^k w_r h_{r,k}(\eta) \right] f_{\mathbf{x}}(\mathbf{x}) d\mathbf{x}. \end{aligned} \quad (25)$$

Thus as  $N \rightarrow \infty$  the total error becomes a function of our ability to model  $g(\eta)$  using a restricted basis set. Throughout the rest of this paper, we will refer to this type of error as the approximation error ( $e_A$ ) or asymptotic error in the proposed method. Now suppose that both  $N$  and  $k$  are restricted to finite values. In this scenario we can no longer assume that  $H_{r,k}(f_0, f_1) = \hat{H}_{r,k}(f_0, f_1)$ , and as a result there is a second type of error that must be accounted for which we will call the estimation error ( $e_{est}$ ). The estimation error can be defined as the difference between the asymptotic value of our linear combination of basis functions and its finite sample estimate

$$e_{est} = \sum_{r=0}^k w_r H_{r,k}(f_0, f_1) - \sum_{r=0}^k w_r \hat{H}_{r,k}(f_0, f_1), \quad (26)$$

and can be thought of as the linear combination of the errors of each individual basis function and their respective weights

$$\begin{aligned} e_{est} &= \sum_{r=0}^k w_r \left[ H_{r,k}(f_0, f_1) - \hat{H}_{r,k}(f_0, f_1) \right] \\ &= \sum_{r=0}^k w_r e_{r,k}. \end{aligned} \quad (27)$$

By manipulating (23)

$$\begin{aligned} e_T &= G(f_0, f_1) - \hat{G}(f_0, f_1) \\ &= G(f_0, f_1) - \sum_{r=0}^k w_r \hat{H}_{r,k}(f_0, f_1) \\ &= G(f_0, f_1) - \sum_{r=0}^k w_r H_{r,k}(f_0, f_1) \\ &\quad + \sum_{r=0}^k w_r H_{r,k}(f_0, f_1) - \sum_{r=0}^k w_r \hat{H}_{r,k}(f_0, f_1) \\ &= e_A + e_{est}, \end{aligned} \quad (28)$$

we can show that the total error is simply the sum of the approximation error and the estimation error. Understanding the trade-off between these two error types is essential.

#### B. Considerations for finite $k$

A finite sample also implies a finite  $k$  and impacts the approximation error. Let us consider the Bernstein weighting scheme introduced in (21) for the scenario where the size of the basis set ( $k$ ) is restricted. Consider the following example problem.

*Example:* Suppose that we wish to estimate the function

$$g(\eta) = \binom{3}{1} \eta (1 - \eta)^2 \quad (29)$$

using the basis set  $\beta_{0,3}(\eta), \beta_{1,3}(\eta), \beta_{2,3}(\eta), \beta_{3,3}(\eta)$ . Because  $g(\eta) = \beta_{1,3}(\eta)$ , there exists a set of weights such that

$$\sum_{r=0}^3 w_r \beta_{r,3}(\eta) = g(\eta), \quad (30)$$

however, using the Bernstein weighting scheme in (21) yields

$$\begin{aligned} \hat{g}(\eta) &= \sum_{r=0}^3 g\left(\frac{r}{3}\right) \beta_{r,3}(\eta) \\ &= g\left(\frac{0}{3}\right) \beta_{0,3}(\eta) + g\left(\frac{1}{3}\right) \beta_{1,3}(\eta) + g\left(\frac{2}{3}\right) \beta_{2,3}(\eta) \\ &\quad + g\left(\frac{3}{3}\right) \beta_{3,3}(\eta) \\ &= \frac{4}{3} \eta (1 - \eta)^2 + \frac{2}{3} \eta^2 (1 - \eta) \\ &\neq g(\eta). \end{aligned} \quad (31)$$

It is clear from this example that the Bernstein weighting procedure do not always provide ideal weights when  $k$  is restricted. Based on these results, we are motivated to explore alternative weighting procedures in order to improve the performance of this method for the finite sample case. In the following Section, we will introduce a method of finding better weights using convex optimization.

The second question we must consider is how to select  $k$ . Theoretically we can choose  $k$  to be as small as 1, meaning we only have two functions in our basis set, or as large as  $N - 1$ ; however when  $k = N - 1$  the neighborhood set  $\mathcal{N}_k(\mathbf{x})$  is the same for all points  $\mathbf{x} \in \mathbf{X}$  and all but one basis function will be equal to zero. It is difficult to imagine a set of circumstances for which either of these extremes will ever be ideal, however there are a range of  $k$ 's between these two extremes that are likely more suitable.

In general, there are two major competing factors that must be considered when selecting  $k$ . The first is that the Weierstrass approximation theorem can exactly represent any generator function  $g(\eta)$  as a linear combination of the proposed basis set only as  $k \rightarrow \infty$ . This provides motivation for choosing a large  $k$ -value to ensure the best possible fit of  $g(\eta)$ . The second factor is that

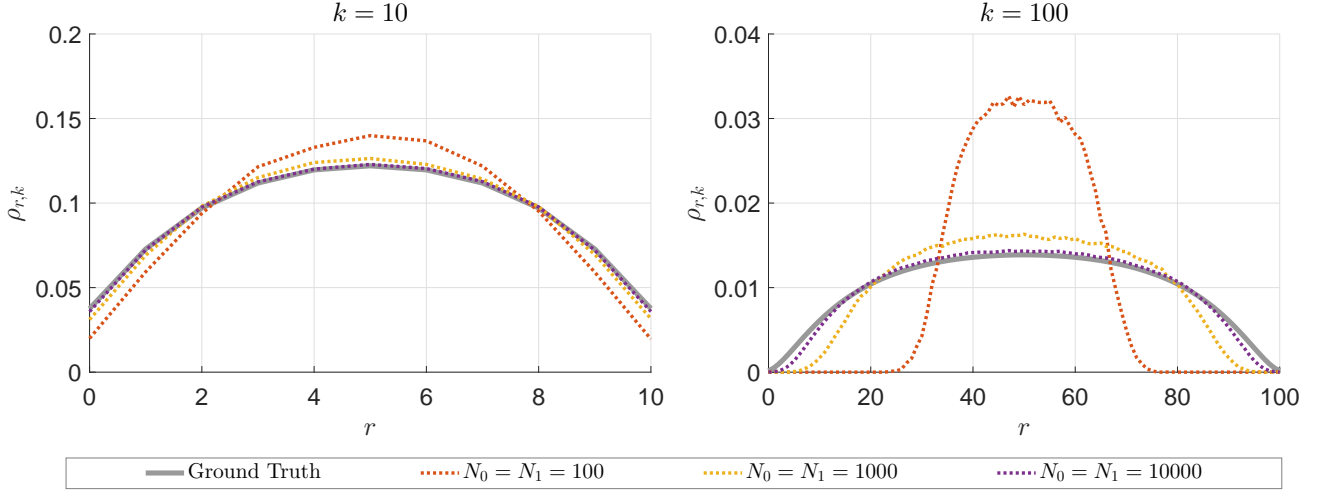


Fig. 2: Plot of true and estimated basis values vs.  $r$  for data drawn from underlying distributions  $f_0(\mathbf{x}) \sim \mathcal{N}(\mathbf{0}_3, \mathbf{I}_3)$  and  $f_1(\mathbf{x}) \sim \mathcal{N}(\frac{1}{\sqrt{3}}\mathbf{1}_3, \mathbf{I}_3)$

the asymptotic characteristics of  $\rho_{r,k}$  are dependent on all points in  $\mathcal{N}_k(\mathbf{x}^*)$ . Moreover the regime for which Theorem 2 holds, requires that  $\frac{k}{N} \rightarrow 0$ , so we are motivated to select  $k$  such that  $N \gg k$ . This means that the selection of  $k$  must achieve a compromise in the trade-off between the approximation and estimation errors, since larger values of  $k$  will increase the amount of finite sample error made in estimating the individual basis functions, while lower values of  $k$  may inhibit our ability to accurately model the desired function in the asymptotic regime.

To illustrate how our ability to estimate the desired set of basis functions varies with  $k$ , we calculate the estimated and true values of  $\rho_{r,k}$  for  $k = 10$  and  $k = 100$  on data drawn from distributions  $f_0(\mathbf{x}) \sim \mathcal{N}(\mathbf{0}_3, \mathbf{I}_3)$  and  $f_1(\mathbf{x}) \sim \mathcal{N}(\frac{1}{\sqrt{3}}\mathbf{1}_3, \mathbf{I}_3)$ , where  $\mathbf{0}_3 = [0 \ 0 \ 0]$  and  $\mathbf{1}_3 = [1 \ 1 \ 1]$ , and plot the results in Figure 2. Estimates are calculated at 3 different sample sizes ( $N_0 = N_1 = 100, 1000, 10000$ ) and each estimate shown in Figure 2 has been averaged across 500 Monte Carlo trials. While we can estimate the basis set for either  $k$ -value with a high degree of accuracy given enough samples, the estimates for  $k = 10$  are noticeably more accurate. In fact, we are able to do about as well with 1000 samples for  $k = 10$  as we are with 10000 samples for  $k = 100$ .

## VI. OPTIMIZATION CRITERIA FOR FITTING DENSITY FUNCTIONALS

In this section, we propose a convex optimization criterion to identify appropriate weights for fitting information-theoretic functions when  $k$  and  $N$  are restricted. Inherently, our goal is to minimize the total error, defined in (23), however minimizing this quantity directly isn't feasible since the value of  $G(f_0, f_1)$  is unknown. To circumvent this challenge we focus on developing a criterion to minimize the approximation error. We initially develop an optimization criterion that assumes the posterior is uniformly distributed, then propose an alternate method which incorporates an estimate of the posterior density function in order to more accurately model the approximation error.

### A. Uniform Optimization Criteria

Recall that the approximation error  $e_A$  can be represented as

$$e_A = \int \epsilon(\eta) f_{\mathbf{x}}(\mathbf{x}) d\mathbf{x}, \quad (32)$$

where

$$\epsilon(\eta) = g(\eta) - \sum_{r=0}^k w_r h_{r,k}(\eta). \quad (33)$$

Since solving (32) requires high-dimensional integration and knowledge of the underlying distributions. However, because  $\eta$  is a function of  $\mathbf{x}$ ,  $\epsilon(\eta)$  is implicitly a function of  $\mathbf{x}$  as well, and by the law of the unconscious statistician [47],

$$e_A = E[\epsilon(\eta)] = \int \epsilon(\eta) f_{\eta}(\eta) d\eta, \quad (34)$$

where  $f_{\eta}(\eta)$  is the probability density function of the random variable  $\eta$ . Rewriting the error in this form simplifies the region of integration to a well defined space (since  $\eta \in [0, 1]$ ) and circumvents the high dimensionality of  $\mathbf{x}$ . While this eliminates some of the challenges in calculating the error it also creates new ones stemming from the fact that  $f_{\eta}(\eta)$  is unknown and difficult to estimate due to its implicit dependency on  $f_0(\mathbf{x})$  and  $f_1(\mathbf{x})$ . The task of estimating  $f_{\eta}(\eta)$  will be explored in detail in Section VI-B, however for the time being we will bypass this challenge and simply attempt to minimize

$$e_A^* = \int |\epsilon(\eta)|^2 d\eta. \quad (35)$$

It is worth noting that if  $f_{\eta}(\eta)$  is uniformly distributed

$$e_A^* = E[|\epsilon(\eta)|^2] \geq e_A^2. \quad (36)$$

If we define a discretized set of posterior values  $\tilde{\eta}_1, \tilde{\eta}_2, \dots, \tilde{\eta}_{\tilde{N}}$ , where  $0 \leq \tilde{\eta}_1 < \tilde{\eta}_2 < \dots < \tilde{\eta}_{\tilde{N}} \leq 1$ , a procedure to identify weights that minimize (35) can be defined as

$$w_0, \dots, w_K = \underset{w_0, \dots, w_K}{\operatorname{argmin}} \frac{1}{\tilde{N}} \sum_{i=1}^{\tilde{N}} \left| g(\tilde{\eta}_i) - \sum_{r=0}^K w_r h_r(\tilde{\eta}_i) \right|^2. \quad (37)$$

To illustrate the effectiveness of this method, we consider the example problem of trying to estimate the Hellinger distance (a problem we will further explore in Section VII). If we assume both classes have equal prior probability ( $p_0 = p_1 = 0.5$ ), then the generator function for the squared Hellinger distance is

$$g(\eta) = (\sqrt{\eta} - \sqrt{1-\eta})^2. \quad (38)$$

This function is estimated using this convex weighting procedure as

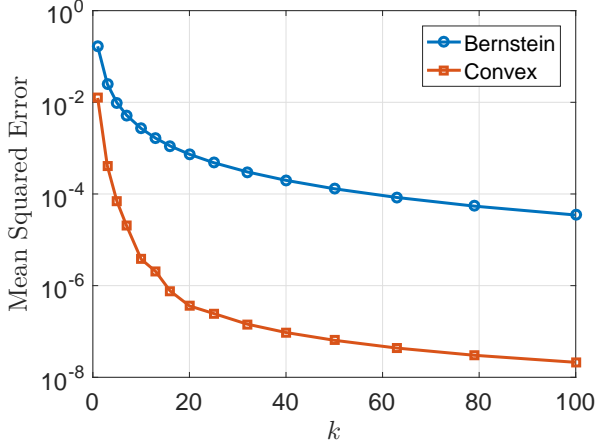


Fig. 3: Approximation error of each fitting procedure as a function of the number of basis elements  $k$ . This is the idealized case where the basis estimation error is zero and the total error is solely due to imperfect approximation of the generator function  $g(\eta)$ .

well as the previously described Bernstein weighting procedure, and we compare how well each method models the desired function for values of  $k$  varying from 0 to 100. The performances of each method is evaluated by the following formula

$$\text{MSE}(\hat{g}, g) = \sum_{i=1}^{\tilde{N}} \left| g(\tilde{\eta}_i) - \hat{g}(\tilde{\eta}_i) \right|^2, \quad (39)$$

and the results are presented in Figure 3 for a range of  $k$  values varying from 1 to 100. This experiment shows that the proposed convex fitting procedure is able to approximate the Hellinger generator function far more accurately than the Bernstein approximation.

The expression (39) does not take into account finite sample errors that lead to noisy estimates of the basis functions and thus does not directly reflect our ability to estimate  $G(f_0, f_1)$  with a finite sample. Additionally, it does not account for the possibility that  $\eta$  is distributed non-uniformly. To empirically examine the finite sample properties of the two approaches, we repeat the previous experiment, this time estimating the basis functions empirically on samples of data drawn from distributions  $f_0(\mathbf{x}) \sim \mathcal{N}(\mathbf{0}_3, \mathbf{I}_3)$  and  $f_1(\mathbf{x}) \sim \mathcal{N}(\frac{1}{\sqrt{3}}, \mathbf{I}_3)$ . We generate  $N = 1000$  samples (500 samples per class) in each of the 500 iterations of a Monte Carlo simulation, and evaluate the MSE as

$$\text{MSE}(G, \hat{G}) = \frac{1}{N_{MC}} \sum_{i=1}^{N_{MC}} [G(f_0, f_1) - \hat{G}(f_0, f_1)]^2, \quad (40)$$

where  $N_{MC}$  represents the number of Monte Carlo iterations. Since we know that the estimation error is scaled by the magnitude of the weights, we also evaluate a modified fitting routine which augments (37) with a regularization term to penalize solutions with large weights,

$$w_0, \dots, w_K = \underset{w_0, \dots, w_K}{\operatorname{argmin}} \frac{1}{\tilde{N}} \sum_{i=1}^{\tilde{N}} \left| g(\tilde{\eta}_i) - \sum_{r=0}^K w_r h_r(\tilde{\eta}_i) \right|^2 + \frac{\lambda}{k} \sum_{r=0}^K w_r^2, \quad (41)$$

where  $\lambda$  represents a tuning parameter which controls the importance assigned to minimization of the approximation error relative to the estimation error. Intuitively, higher  $\lambda$  values will make sense for

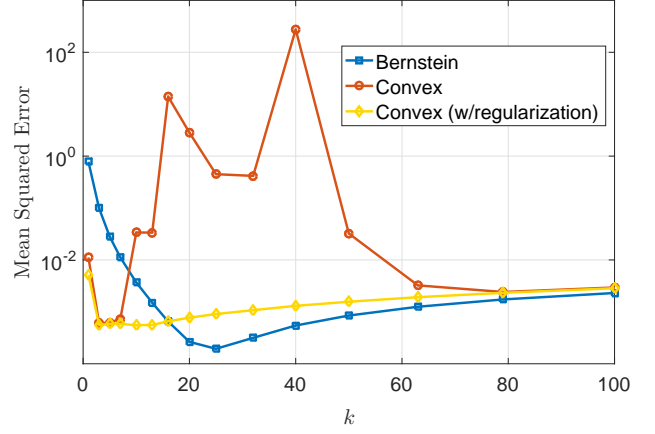


Fig. 4: The total error of each fitting procedure as a function of  $k$ , when there is both approximation and estimation error.

smaller data sets to control the variance of the estimator. We set  $\lambda = 0.01$  for all experiments conducted in this paper. The results of this experiment are shown in Figure 4. We immediately see the necessity of the regularization term, as without it the error becomes extremely large for a range of  $k$  values. More generally, the inclusion of the regularization term improves the performance at every  $k$  value in this experiment. In comparing the Bernstein weights with the convex (regularized) weights, we find that 1) the performance of the convex method is less dependent on the selection of  $k$  and 2) the convex weights generally perform better at lower values of  $k$ , while the Bernstein weights outperform at higher  $k$ . While the peak performance of the Bernstein method is higher than the convex method, there exists no good method of selecting  $k$  *a priori* in order to reliably achieve this performance. In contrast, the convex method with regularization is less sensitive to the value of  $k$  selected.

### B. Density-weighted Optimization Criteria

In the optimization criteria in the previous section we implicitly make the assumption that the distribution of the random variable  $\eta(\mathbf{x}) \sim f_\eta(\eta)$  is uniformly distributed. In this section we will investigate a data-driven estimator for  $f_\eta(\eta)$ , however before we proceed it is important to clarify what this distribution actually is.

We initially introduced  $\eta$  as the posterior likelihood function for class 1, which we showed in (2) can be represented as a function of the underlying distributions. When this function's input is a known point  $\mathbf{x}^*$ ,  $\eta(\mathbf{x}^*)$  represents the probability that  $y = 1$  given that  $\mathbf{x} = \mathbf{x}^*$ . However, if the input is a random variable  $\mathbf{x}$ , then  $\eta$  is also a random variable, which is distributed according to  $f_\eta(\eta)$ .

Figure 5 illustrates  $f_\eta(\eta)$  for two univariate normal distributions  $f_0(x) \sim \mathcal{N}(0, 1)$  and  $f_1(x) \sim \mathcal{N}(1, 1)$ . Figure 5a displays the two class distributions across  $x$ , Figure 5b displays  $\eta(x)$  as a function of  $x$ , and Figure 5c displays  $f_\eta(\eta)$  as a function of  $\eta$ . We see from this illustration that while,  $\eta(x)$  is close to 0 or 1 across most of the region of  $x$  that is displayed,  $\eta(x)$  remains close to 0.5 in the regions where  $f_x(x)$  is greatest. As a result  $f_\eta(\eta)$  is roughly bell-shaped, and the likelihood of  $\eta$  existing at the extremities (close to 0 or 1) is relatively low. Because the probability density in these regions is low, the accuracy of our fit in these regions is less important, and can be given less weight in the fitting routine. Figure 6 repeats this illustration for two well separated normal distributions. In this case the distribution of  $\eta$  is such that  $f_\eta(\eta)$  is most dense towards the extremities, and therefore they should be given more weight in the

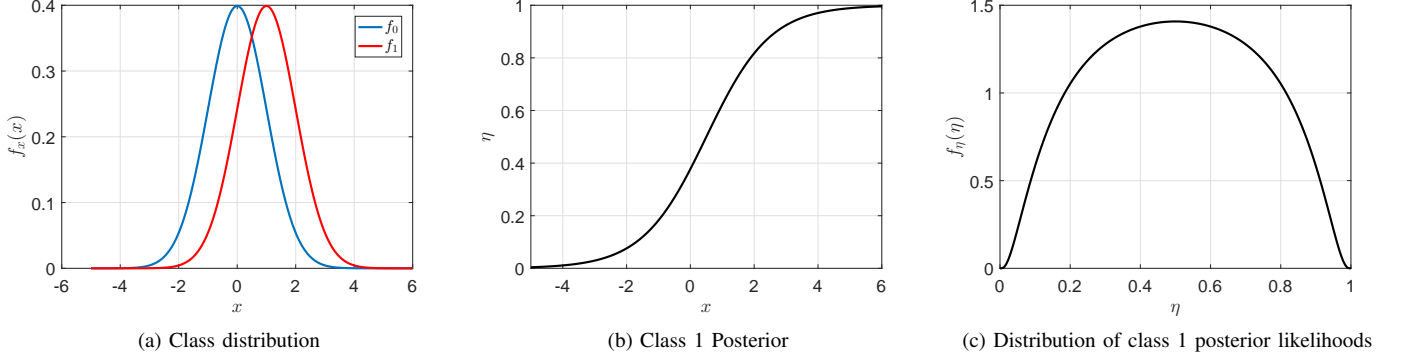


Fig. 5: Illustration of the posterior distribution for two close univariate normal distributions

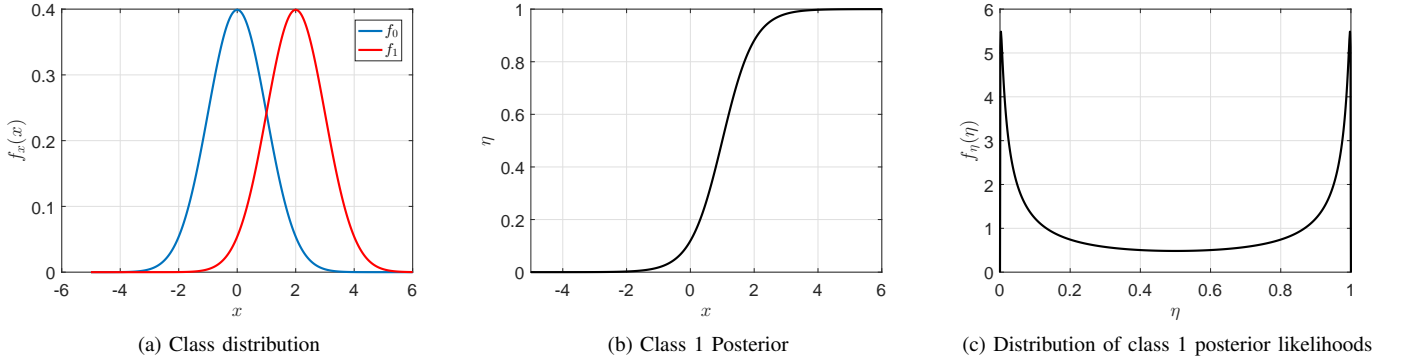


Fig. 6: Illustration of the posterior distribution for two separated univariate normal distributions

fitting routine. Side by side, these two figures present an interesting contrast. Despite the fact that the set of density functions  $f_0$  and  $f_1$  look quite similar in the two scenarios, the minor difference in separation significantly alters the distribution of  $\eta$ .

In practice the underlying distributions  $f_0$  and  $f_1$  are unknown, and as a result,  $f_\eta(\eta)$  is unknown as well. However, if we were able to sample  $f_\eta(\eta)$  at  $\eta = \tilde{\eta}_i$ , a more direct method of minimizing  $e_T$  would be to solve

$$w_0, \dots, w_K = \underset{w_0, \dots, w_K}{\operatorname{argmin}} \sum_{i=1}^{\tilde{N}} \left| g(\tilde{\eta}_i) - \sum_{r=0}^K w_r h_r(\tilde{\eta}_i) \right|^2 \hat{f}_\eta(\tilde{\eta}_i) \Delta_{\tilde{\eta}} + \frac{\lambda}{k} \sum_{r=0}^K w_r^2, \quad (42)$$

where  $\Delta_{\tilde{\eta}} = \tilde{\eta}_{i+1} - \tilde{\eta}_i$ .

Below we show that  $\rho_{r,k}$ , the statistic previously defined in Theorem 1 can be used to sample the PDF of  $\eta$ . This result is stated in Theorem 3.

**Theorem 3.** As  $N \rightarrow \infty$  and  $k \rightarrow \infty$  in a linked manner such that  $\frac{k}{N} \rightarrow 0$

$$f_\eta\left(\frac{r}{k}\right) = k\rho_{r,k}$$

almost surely.

Theorem 3 is useful as it provides a method of sampling  $f_\eta(\eta)$  that doesn't depend on estimates of the underlying density functions  $f_0$  and  $f_1$ . Using this result, we can estimate the density of the posterior at point  $\tilde{\eta}_i$  as

$$\hat{f}_\eta(\tilde{\eta}_i) = \tilde{k}_i \rho_{\tilde{r}_i, \tilde{k}_i} \quad (43)$$

where  $\tilde{\eta}_i = \frac{\tilde{r}_i}{\tilde{k}_i}$ . Sampling at exactly  $\tilde{\eta}_i$  may not always be possible since the maximum value of  $k$  is limited by the sample size, and  $k$  determines the resolution of the sampling scheme. Even if it is possible, it may not be desirable to recalculate  $\rho_{\tilde{r}_i, \tilde{k}_i}$  for different values of  $k_i$  because of the computational burden. To overcome these problems we can design our approach such that we utilize the same set of test statistics  $\rho_{r,k}$  in the estimation of the posterior distribution as are used in the estimation of the basis set. One way to do this is to assign the set of discretized posteriors

$$\tilde{\eta}_1, \tilde{\eta}_2, \tilde{\eta}_3, \dots, \tilde{\eta}_{\tilde{N}} = 0, \frac{1}{k}, \frac{2}{k}, \dots, 1 \quad (44)$$

so that it is straightforward to calculate  $\hat{f}_\eta(\eta)$  from the known values of  $\rho_{r,k}$ . An alternate approach is to leave  $\tilde{\eta}$  unchanged and interpolate  $\hat{f}_\eta(\eta)$  to determine its value at the desired points. The advantage of this approach is that it doesn't constrain how  $\eta$  is sampled. Throughout the rest of this paper, we will employ the latter method and solve for  $\hat{f}_\eta(\eta)$  by linearly interpolating between the already known values.

Because this density-weighted fitting routine more directly minimizes the approximation error of the final estimate, we expect it to generally outperform the uniform method, particularly in cases where the density of the posterior is highly non-uniform and where the desired generator function  $g(\eta)$  is difficult to model using the proposed basis set. This hypothesis will be verified in Sections VII and VIII, when we empirically evaluate our methods with real data.



## VII. DIVERGENCE ESTIMATION

In this section, we show how the proposed method can be applied to estimating three  $f$ -divergences, the Hellinger distance, the KL-divergence, and the  $D_p$ -divergence, directly from data.

### A. Hellinger Distance

The Hellinger distance squared is an  $f$ -divergence used to quantify the dissimilarity between two probability distributions and is calculated by

$$H^2(f_0, f_1) = \frac{1}{2} \int \left( \sqrt{f_0(\mathbf{x})} - \sqrt{f_1(\mathbf{x})} \right)^2 d\mathbf{x}. \quad (45)$$

Using the approach proposed in Section VI, we can estimate  $H^2(f_0, f_1)$  by fitting weights to the generator function

$$g_H(\eta) = \frac{1}{2} \left( \sqrt{\frac{\eta}{p_1}} - \sqrt{\frac{1-\eta}{p_0}} \right)^2. \quad (46)$$

To evaluate the efficacy of this method, we conduct four different experiments in which we attempt to estimate the Hellinger distance between two distributions from finite sample data. In the first three experiments, both distributions are normally distributed according to  $f(\mathbf{x}) \sim N(\mu \mathbf{1}_d, \Sigma_d)$ , where

$$\Sigma_d = \begin{bmatrix} \sigma_{1,1} & \sigma_{1,2} & \dots & \sigma_{1,d} \\ \sigma_{2,1} & \sigma_{2,2} & \dots & \sigma_{2,d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{d,1} & \sigma_{d,2} & \dots & \sigma_{d,d} \end{bmatrix} \quad (47)$$

for  $\sigma_{i,j} = \beta^{|i-j|}$ . The first experiment evaluates the most basic case where both Gaussians have spherical covariance. The second experiment considers the case where there exists a strong fixed dependency between adjacent dimensions by using an elliptical covariance structure. The third experiment evaluates the case where this dependency between adjacent dimensions is now dependent on which class the data is drawn from. In the fourth experiment, we return to linearly independent dimensions and consider the case where one of the distributions isn't normally distributed, but instead uniformly distributed according to

$$f(\mathbf{x}) = \begin{cases} \frac{1}{(2\beta)^d} & \mu - \beta \leq x \leq \mu + \beta \quad \forall x \in \mathbf{x} \\ 0 & \text{otherwise.} \end{cases} \quad (48)$$

A detailed description of the distribution types and parameter setting used in each of the four experiments is presented in Table I. In addition to using the proposed method, we also estimate the Hellinger distance using two different plug-in estimators, one based on a parametric density estimator that assumes the data is normally distributed and one based on a  $k$ -NN density estimate of the underlying distributions. To calculate the  $k$ -NN estimate, we use the universal divergence estimation approach described in [48] and implemented in the ITE toolbox [49]. This method allows us to fix  $k = 10$  and still achieve an asymptotically consistent estimate of the divergence.

In each of the first three experiments, the parametric model shows the highest rate of convergence as expected, although in experiment two it is slightly outperformed at smaller sample sizes by the proposed method. In the fourth experiment, when the assumption of Gaussianity in  $f_1$  no longer holds, the parametric solution is significantly biased and as a result, is outperformed by both non-parametric methods at higher sample sizes ( $N > 2000$ ). Relative to the  $k$ -NN plug-in estimator, the proposed method performs slightly worse in experiment 1, slightly better in experiment 2 and significantly better in experiments 3 and 4, with the results

TABLE I: Experiment overview table

	$f_0(\mathbf{x})$			$f_1(\mathbf{x})$		
	Family	$\mu$	$\beta$	Family	$\mu$	$\beta$
Experiment 1	Normal	0	0	Normal	$\sqrt{\frac{1}{3}}$	0
Experiment 2	Normal	0	0.8	Normal	$\sqrt{\frac{1}{3}}$	0.8
Experiment 3	Normal	0	0.8	Normal	$\sqrt{\frac{1}{3}}$	0.9
Experiment 4	Normal	0	0	Uniform	0	3

being relatively consistent across the various sample sizes. The improvement in performance shown in experiments 3 and 4 suggests that the proposed method offers the greatest benefit when there exists differences in the shapes of the two underlying distributions. Though the density-weighted procedure consistently outperformed the uniform method, the observed improvement was relatively minor in these experiments.

### B. Kullback Leibler Divergence

The Kullback Leibler (KL) divergence [50], also sometimes referred to as the KL risk or relative entropy, is an asymmetric measure of divergence between two probability density functions. Using our regular notation the KL-divergence can be calculated by

$$d_{KL}(f_0||f_1) = \int_{-\infty}^{\infty} f_0(\mathbf{x}) \log \left( \frac{f_0(\mathbf{x})}{f_1(\mathbf{x})} \right) d\mathbf{x}. \quad (49)$$

While the KL-divergence has the same general purpose as the Hellinger distance, that is to measure the dissimilarity between two probability density functions, it also has several key difference. Firstly since the KL-divergence is asymmetric it doesn't technically qualify as a distance function and  $d_{KL}(f_0||f_1)$  isn't necessarily equal to  $d_{KL}(f_1||f_0)$ . This also means that  $d_{KL}(f_0||f_1)$  and  $d_{KL}(f_1||f_0)$  will have different generator functions. We can define the generator function for  $d_{KL}(f_0||f_1)$  as

$$g_{KL}^0(\eta) = \frac{1-\eta}{p_0} \log \left( \frac{p_1(1-\eta)}{p_0\eta} \right) \quad (50)$$

and the generator function for  $d_{KL}(f_1||f_0)$  as

$$g_{KL}^1(\eta) = \frac{\eta}{p_1} \log \left( \frac{p_0\eta}{p_1(1-\eta)} \right) \quad (51)$$

such that

$$d_{KL}(f_i||f_{i-1}) = \int_{-\infty}^{\infty} g_{KL}^i(\eta) (p_0 f_0(\mathbf{x}) + p_1 f_1(\mathbf{x})) d\mathbf{x}. \quad (52)$$

It is worth noting that when  $p_0 = p_1 = 0.5$

$$g_{KL}^1(\eta) = g_{KL}^0(1-\eta) \quad (53)$$

thus  $g_{KL}^1(\eta)$  is a reflection of  $g_{KL}^0(\eta)$ . One challenge presented in modeling the KL-divergence is that the generator functions are difficult to model at the end points, since  $g_{KL}^0(0) = \infty$  and  $g_{KL}^0(1)$  is undefined though

$$\lim_{\eta \rightarrow 1^-} g_{KL}^0(\eta) = 0. \quad (54)$$

Due to their symmetry  $g_{KL}^1$  has the same problem at the opposite endpoints. To handle this we simply select our discretized set of posteriors  $\tilde{\eta}_1, \tilde{\eta}_2, \dots, \tilde{\eta}_{\tilde{N}}$  such that  $0 < \tilde{\eta}_1 < \tilde{\eta}_2 < \dots < \tilde{\eta}_{\tilde{N}} < 1$ . For the experiments in this Section, we set

$$\tilde{\eta}_1, \tilde{\eta}_2, \tilde{\eta}_3, \dots, \tilde{\eta}_{100}, \tilde{\eta}_{101} = \epsilon, 0.01, 0.02, \dots, 0.99, 1 - \epsilon \quad (55)$$

where  $\epsilon = 10^{-4}$ . Using this modified set of discretized posteriors, we repeat the set of four experiments described in Section VII-A to



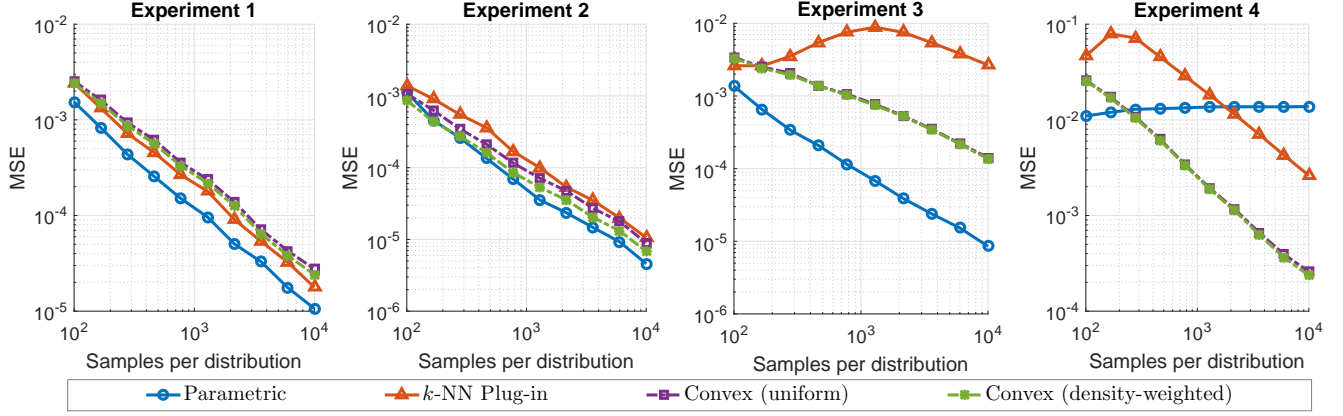


Fig. 7: Plots of MSE vs. Sample size in estimating the Hellinger distance for the four different experiments outlined in Table I

evaluate the proposed methods ability to estimate the KL-divergence. The results of this experiment are displayed in Figure 8. Like the estimates of the Hellinger distance, the parametric method generally yielded the best performance in the first three experiments, but suffered from a large asymptotic bias in experiment 4. The proposed method once again significantly outperformed the  $k$ -NN plug-in estimator in experiments 3 and 4, however the results in experiments 1 and 2 are slightly more nuanced due to the significant difference in performance between the two optimization criteria in these trials. In both of these trials the density-weighted criteria significantly outperforms the uniform method at all sample sizes. In experiment 1 the  $k$ -NN plug-in estimator outperforms the regular plug-in estimator at all sample sizes, but is outperformed by the density-weighted method for  $N > 300$ . In experiment 2 the  $k$ -NN plug-in estimator consistently outperforms both proposed methods, however the improvement over the density-weighted method is marginal.

### C. $D_p$ -Divergence

The  $D_p$ -divergence is an  $f$ -divergence defined by

$$D_{p_0}(f_0, f_1) = \frac{1}{4p_0p_1} \left[ \int \frac{(p_0f_0(\mathbf{x}) - p_1f_1(\mathbf{x}))^2}{p_0f_0(\mathbf{x}) + p_1f_1(\mathbf{x})} d\mathbf{x} - (p_0 - p_1)^2 \right]. \quad (56)$$

The  $D_p$ -divergence has the unique property of being directly estimable from data using minimum spanning trees [12]. Because of this property, it has been used to form non-parametric estimates of the Fisher information [37] as well as upper bounds on the Bayes error rate in a range of classification problems [12], [51]. The generator function for the  $D_p$ -divergence can be defined as

$$g_{D_p}(\eta) = \frac{(2\eta - 1)^2 - (2p_0 - 1)^2}{4p_0(1 - p_0)} \quad (57)$$

which simplifies to  $(2\eta - 1)^2$  when  $p_0 = p_1 = 0.5$ . We once again repeat the experiments described in Section VII-A to evaluate the proposed methods ability to estimate the  $D_p$ -divergence. This experiment provides the unique opportunity to compare the proposed method to a more traditional direct estimation procedure. The results of this experiment are displayed in Figure 9.

As in the previous experiments, the parametric estimate generally performed the best in the first three experiments, but suffered from a large asymptotic bias in experiment 4. The proposed methods perform better than the MST-based estimation in experiments 1 and 2 but worse in experiments 3 and 4. The relative performance of each method in this experiment was largely consistent across the various sample sizes, though the proposed method seems to be converging

slightly faster than the MST method in experiment 4 and could possibly exceed its performance given enough samples. Unlike in the previous experiments, we found no difference in performance between the uniform optimization criteria and the density-weighted criteria in this experiment. This is due to the fact that  $g_{D_p}(\eta)$  is a polynomial and can be perfectly represented by the proposed basis set, even when  $k$  is truncated. Since we are able to achieve a solution where  $\epsilon(\eta) = 0 \forall \eta$ , the density weighting has no impact on the final results.

### VIII. FITTING BOUNDS ON PERFORMANCE

The optimization criteria in the proposed fitting routines gives us the ability to not only approximate information-theoretic functions, but to bound them as well. This is especially useful for forming bounds on the Bayes Error Rate (BER). The Bayes error rate represents the optimal classification performance that is achievable for a given pair of class distributions  $f_0(\mathbf{x})$  and  $f_1(\mathbf{x})$  with prior probabilities  $p_0$  and  $p_1$  respectively and can be calculated by

$$\epsilon^{\text{Bayes}} = \int_{p_0f_0(\mathbf{x}) \leq p_1f_1(\mathbf{x})} p_0f_0(\mathbf{x})d\mathbf{x} + \int_{p_1f_1(\mathbf{x}) \leq p_0f_0(\mathbf{x})} p_1f_1(\mathbf{x})d\mathbf{x}. \quad (58)$$

In essence the BER measures the intrinsic difficulty of a particular classification problem based on the data. A thorough understanding of the BER of a particular problem can help design optimal classifiers. Because of the challenges associated with estimating the BER, much of the literature has focused on generating bounds on the BER [12], [34], which are generally formulated in terms of some measure of divergence between the two class distributions. One such bound, the well-known Bhattacharyya bound, is given by [34]

$$\frac{1}{2} - \frac{1}{2}\sqrt{1 - BC^2(f_0, f_1)} \leq \epsilon^{\text{Bayes}} \leq \frac{1}{2}BC(f_0, f_1), \quad (59)$$

where

$$BC(f_0, f_1) = 1 - H^2(f_0, f_1). \quad (60)$$

While the Hellinger distance here can be estimated via any of the methods discussed in Section VII-A, parametric estimates are most common. Alternatively [12] introduced the bounds

$$\frac{1}{2} - \frac{1}{2}\sqrt{D_{\frac{1}{2}}(f_0, f_1)} \leq \epsilon^{\text{Bayes}} \leq \frac{1}{2} - \frac{1}{2}D_{\frac{1}{2}}(f_0, f_1) \quad (61)$$

where

$$D_{\frac{1}{2}}(f_0, f_1) = 1 - 2 \int \frac{f_0(\mathbf{x})f_1(\mathbf{x})}{f_0(\mathbf{x}) + f_1(\mathbf{x})} d\mathbf{x}. \quad (62)$$

These bounds have the advantage of being provably tighter than the Bhattacharyya bounds [12]. Furthermore since  $D_{\frac{1}{2}}$  represents

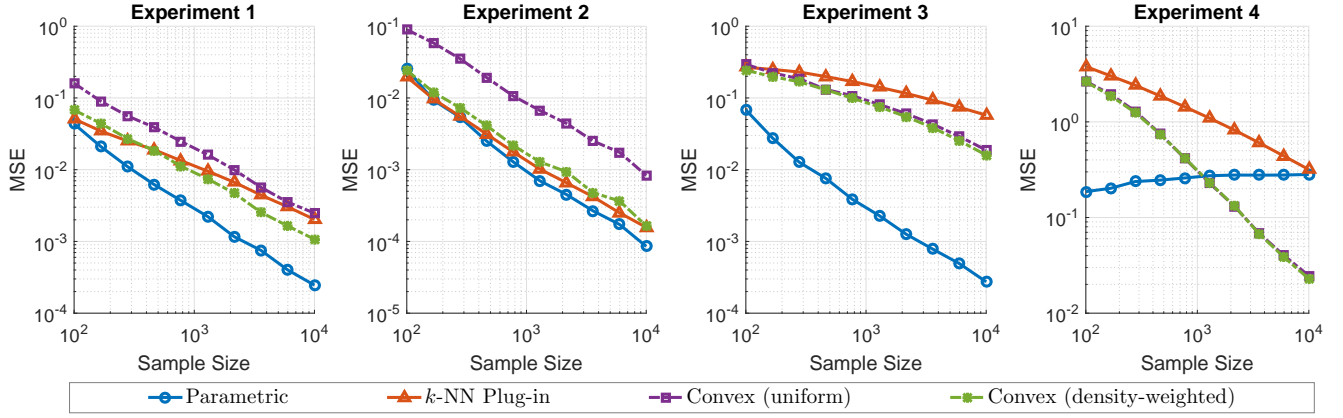


Fig. 8: Plots of MSE vs. Sample size in estimating the KL-divergence for the four different experiments outlined in Table I

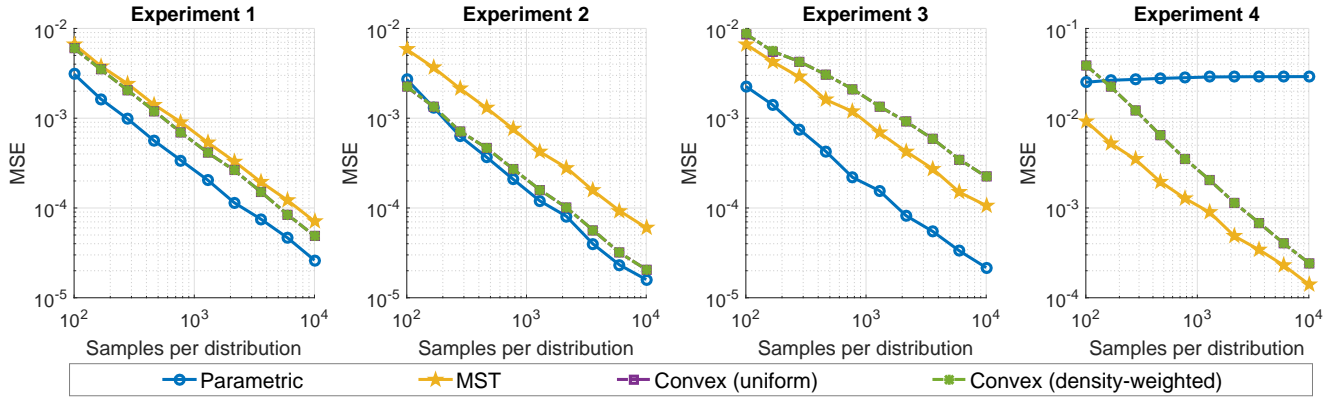


Fig. 9: Plots of MSE vs. Sample size in estimating the  $D_p$ -divergence for the four different experiments outlined in Table I

a particular case of the  $D_p$ -divergence, which is estimable directly from data, these bounds bypass the need for density estimation much like the approaches proposed in this paper. While these bounds are significantly tighter than the Bhattacharyya bounds, they still leave room for improvement. Avi-Itzhak proposed arbitrarily tight bounds on the BER in [36], however these bounds require density estimation to be employed in practical problems. In this section, we will use a modified version of the previously described fitting routine in order to investigate how tightly we are able to bound the BER using a linear combination of directly estimable basis functions.

Using the fitting routine described in (41) to bound the BER, requires that we define  $g(\eta)$  appropriately for estimation of the BER, and constrain our fit such that

$$\sum_{r=0}^K w_r h_r(\tilde{\eta}_i) \geq g(\tilde{\eta}_i) \quad \forall \tilde{\eta}_i. \quad (63)$$

We can express (58) as

$$\begin{aligned} \epsilon^{\text{Bayes}} &= \int \min [p_0 f_0(\mathbf{x}), p_1 f_1(\mathbf{x})] d\mathbf{x} \\ &= \int \min [1 - \eta, \eta] f_{\mathbf{x}}(\mathbf{x}) d\mathbf{x} \end{aligned} \quad (64)$$

so  $g(\eta) = \min [1 - \eta, \eta]$ . Incorporating these changes within the

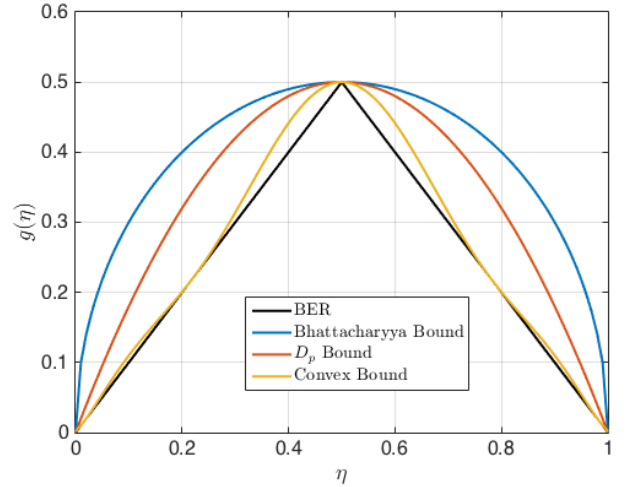


Fig. 10: The Bayes error rate along with the three considered upper bounds displayed as a function of  $\eta$ .

regularized fitting routine described in (41) yields

$$\begin{aligned} w_0, \dots, w_K &= \\ \argmin_{w_0, \dots, w_K} & \frac{1}{\tilde{N}} \sum_{i=1}^{\tilde{N}} \left| g(\tilde{\eta}_i) - \sum_{r=0}^K w_r h_r(\tilde{\eta}_i) \right|^2 + \frac{\lambda}{k} \sum_{r=0}^K w_r^2 \quad (65) \\ \text{subject to} & \sum_{r=0}^K w_r h_r(\tilde{\eta}_i) \geq g(\tilde{\eta}_i) \quad \forall \tilde{\eta}_i. \end{aligned}$$

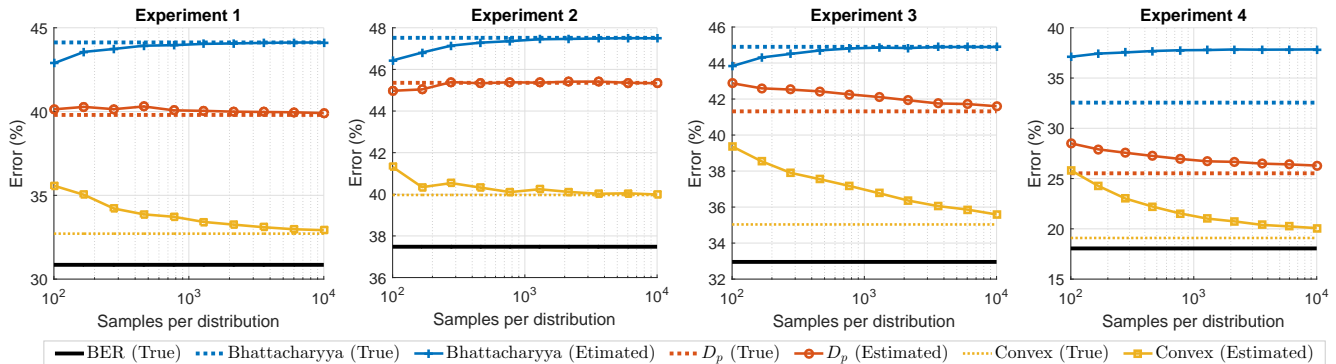


Fig. 11: Plots of theoretical and estimate upper bounds on the BER as a function of sample size for the four different experiments outlined in Table I

Figure 10 compares the theoretical values of each of these bounds as a function of  $\eta$ . These results indicate that the proposed method offers much tighter theoretical bounds than the other two methods, however this bound is based on the asymptotic properties of the proposed basis set and doesn't consider the limitations of a finite sample estimate.

We evaluate the finite-sample performance of this method by calculating each of the described bounds across the four experiments described in Table I. Figure 11 displays the ground truth value of the BER, along with the theoretical and estimated values for each of the three bounds (Bhattacharyya,  $D_p$ , and convex) across sample sizes ranging from 100 to 10000. The Bhattacharyya bound is calculated based on a parametric plug-in estimator which assumes both class distributions to be normally distributed. The  $D_p$  bound is calculated from the Friedman-Rafsky test statistic using the approach described in [12]. Finally the convex bound is calculated as a linear combination of the proposed directly estimable basis functions using weights optimized according to (65). The results of this experiment are largely consistent across the four experiments, the convex method yields the tightest bound, followed by the  $D_p$  bound, and finally the Bhattacharyya bound. Except for the Bhattacharyya bound in experiment 4, which is estimated parametrically, all of the bounds appear to converge to their asymptotic solution. While the convex bound generally offers a slightly slower convergence rate than the other two solutions, it remains tighter than the other two bounds across all sample sizes.

In order to further validate this bound we repeat one of the experiments conducted in [12] by evaluating the proposed bound along with the Mahalanobis bound, the Bhattacharyya bound, and the  $D_p$  bound on two 8-dimensional gaussian data sets described in [14]. The mean and standard deviations of  $f_0$  and  $f_1$  for the two data sets are described in Table II, and all dimensions are independent. These data sets allow us to analyze the tightness and validity of the bounds in a higher dimensional setting. For this experiment the sample size was fixed at  $N = 1000$  and only the empirical value of each of the bounds was evaluated. Table III displays the mean and standard deviation of each bound calculated across 500 Monte Carlo iterations for each of the two data sets. In both data sets the convex method provides the tightest bounds on the BER.

## IX. CONCLUSION

This paper introduces a novel method for estimating density functionals which utilizes a set of directly estimable basis functions. The most appealing feature of the proposed method is its flexibility. Where previous methods of direct estimation are generally only

TABLE II: Parameters for 2 8-dimensional Gaussian data sets for which the Bayes error rate is known (from [52])

$\mathcal{D}_0$	$\mu_1$	0	0	0	0	0	0	0
	$\sigma_0$	1	1	1	1	1	1	1
	$\mu_1$	2.56	0	0	0	0	0	0
	$\sigma_1$	1	1	1	1	1	1	1
$\mathcal{D}_2$	$\mu_0$	0	0	0	0	0	0	0
	$\sigma_0$	1	1	1	1	1	1	1
	$\mu_1$	3.86	3.10	0.84	0.84	1.64	1.08	0.26
	$\sigma_1$	8.41	12.06	0.12	0.22	1.49	1.77	0.35

TABLE III: Comparing upper bounds on the Bayes error rate for the multivariate Gaussians defined in Table II.

	Data 1	Data 2
Actual Bayes Error	10%	1.90%
Mahalanobis Bound	18.90% $\pm$ 0.55%	14.07% $\pm$ 0.45%
Bhattacharyya Bound	21.74% $\pm$ 0.87%	4.68% $\pm$ 0.27%
$D_p$ Bound	16.51% $\pm$ 1.07%	3.99% $\pm$ 0.52%
Convex Bound	<b>14.17% <math>\pm</math> 0.86%</b>	<b>3.87% <math>\pm</math> 0.43%</b>

applicable to a specific quantity, we show that the basis set can be used to generate an asymptotically consistent estimate of a broad class of density functionals, including all  $f$ -divergences and the Bayes error rate. We validate these findings by experimentally evaluating the proposed method's ability to estimate three different divergences (the KL-divergence, the Hellinger distance, and the  $D_p$ -divergence) for four pairs of multivariate probability density functions. The results reveal that the proposed method performs competitively with other non-parametric divergence estimation methods, and seems to outperform them in cases where the data from the two distributions have different covariance structures or belong to different families. Additionally we demonstrate how the method can be modified to generate empirically-estimable bounds on the Bayes error rate that are much tighter than existing bounds.

Future work should focus on studying the finite-sample properties of the basis set proposed in this paper, since this represents a major source of error for the proposed methodology. An improved understanding of these properties could enable us to refine the regularization term in our optimization criteria to more accurately model each weights contribution to the estimation error or to develop ensemble methods, like those in [26], for estimating the individual basis functions.

## APPENDIX A PROOF OF THEOREM 1

Conditioned on  $\mathbf{x}_i$ , we model the binary label,  $y_i$ , as a Bernoulli random variable with unknown probability,  $p_i = \eta(\mathbf{x}_i) = P[y = 1 | \mathbf{x} = \mathbf{x}_i]$ . The probability that  $r$  of the  $k$  instances in  $\mathcal{N}_k(\mathbf{x}^*)$  belong to class 1 (are drawn from  $f_1(\mathbf{x})$ ) can be written in terms of the sum of  $y_i$ 's in the neighborhood  $\mathcal{N}_k(\mathbf{x}^*)$ ,

$$P\left[\sum_{i:\mathbf{x}_i \in \mathcal{N}_k(\mathbf{x}^*)} y_i = r\right]. \quad (66)$$

Each of the Bernoulli random variables is associated with different probability since  $p_i = P[y = 1 | \mathbf{x} = \mathbf{x}_i]$ . In the asymptotic regime (when  $N \rightarrow \infty$ ) all instances  $\mathbf{x} \in \mathcal{N}_k(\mathbf{x}^*)$  converge to a single point,  $\mathbf{x}_i \rightarrow \mathbf{x}^*$  [53]. This means that the corresponding  $y_i$  values become i.i.d. random variables and their sum can be modeled by the binomial distribution. This allows us to express the probability that  $\Phi_k(\mathbf{x}^*) = r$  as

$$P[\Phi_k(\mathbf{x}) = r | \mathbf{x} = \mathbf{x}^*] = \binom{k}{r} \eta^r(\mathbf{x}^*) (1 - \eta(\mathbf{x}^*))^{k-r}. \quad (67)$$

The marginal probability distribution of  $\Phi_k(\mathbf{x})$  is simply found by integration of (67) against  $f_{\mathbf{x}}(\mathbf{x})$ :

$$P[\Phi_k(\mathbf{x}) = r] = \int \binom{k}{r} \eta^r(\mathbf{x}) (1 - \eta(\mathbf{x}))^{k-r} f_{\mathbf{x}}(\mathbf{x}) d\mathbf{x}. \quad (68)$$

We can empirically estimate this quantity using the estimator since, by Borel's law of large numbers,

$$\lim_{N \rightarrow \infty} \rho_{r,k} = P[\Phi_k(\mathbf{x}) = r] \quad (69)$$

therefore

$$\lim_{N \rightarrow \infty} \rho_{r,k} = \int \binom{k}{r} \eta^r (1 - \eta)^{k-r} f_{\mathbf{x}}(\mathbf{x}) d\mathbf{x}. \quad (70)$$

## APPENDIX B PROOF OF THEOREM 2

Starting with the expression

$$\lim_{\substack{N \rightarrow \infty, k \rightarrow \infty \\ k/N \rightarrow 0}} \sum_{r=0}^k g\left(\frac{r}{k}\right) \rho_{r,k} = \lim_{k \rightarrow \infty} \sum_{r=0}^k g\left(\frac{r}{k}\right) \lim_{\substack{N \rightarrow \infty \\ \frac{k}{N} \rightarrow 0}} \rho_{r,k} \quad (71)$$

Using Theorem 1, this can be rewritten as

$$\begin{aligned} & \lim_{k \rightarrow \infty} \sum_{r=0}^k g\left(\frac{r}{k}\right) \int \binom{k}{r} \eta^r (1 - \eta)^{k-r} f_{\mathbf{x}}(\mathbf{x}) d\mathbf{x} \\ & \int \left[ \lim_{k \rightarrow \infty} \sum_{r=0}^k g\left(\frac{r}{k}\right) \binom{k}{r} \eta^r (1 - \eta)^{k-r} \right] f_{\mathbf{x}}(\mathbf{x}) d\mathbf{x}, \end{aligned} \quad (72)$$

which according to Weierstrass' Approximation Theorem simplifies to

$$\begin{aligned} & = \int g(\eta) f_{\mathbf{x}}(\mathbf{x}) d\mathbf{x} \\ & = G(f_0, f_1). \end{aligned} \quad (73)$$

## APPENDIX C PROOF OF THEOREM 3

Consider sample instance  $\mathbf{x}^*$ . First recall that

$$\Phi_k(\mathbf{x}^*) = \sum_{i:\mathbf{x}_i \in \mathcal{N}_k(\mathbf{x}^*)} y_i$$

Note that when  $N \rightarrow \infty$

$$\|\mathbf{x}_i - \mathbf{x}^*\| \rightarrow 0 \quad \forall \mathbf{x}_i \in \mathcal{N}_k(\mathbf{x}^*). \quad (74)$$

and as a result the corresponding  $y_i$  are Bernoulli i.i.d. random variables. Therefore  $\frac{1}{k} \Phi_k(\mathbf{x}^*)$  represents the arithmetic mean of these output values, and

$$\lim_{\substack{N \rightarrow \infty, k \rightarrow \infty \\ k/N \rightarrow 0}} \frac{1}{k} \Phi_k(\mathbf{x}^*) = \lim_{\substack{N \rightarrow \infty, k \rightarrow \infty \\ k/N \rightarrow 0}} \frac{1}{k} \sum_{i:\mathbf{x}_i \in \mathcal{N}_k(\mathbf{x}^*)} y_i \rightarrow \eta. \quad (75)$$

Now, since  $\Phi_k(\mathbf{x})$  must be an integer, we can rewrite (69) as

$$\lim_{\substack{N \rightarrow \infty, k \rightarrow \infty \\ k/N \rightarrow 0}} \rho_{r,k} = P\left[r - 1 < \Phi_k(\mathbf{x}) \leq r\right] \quad (76)$$

and using (75) form

$$\begin{aligned} \lim_{\substack{N \rightarrow \infty, k \rightarrow \infty \\ k/N \rightarrow 0}} \rho_{r,k} &= P\left[\frac{r-1}{k} < \frac{1}{k} \Phi_k(\mathbf{x}) \leq \frac{r}{k}\right] \\ &= P\left[\frac{r-1}{k} < \eta \leq \frac{r}{k}\right] \\ &= F_{\eta}\left(\frac{r}{k}\right) - F_{\eta}\left(\frac{r-1}{k}\right). \end{aligned} \quad (77)$$

Now if we multiply each side by  $k$ , the right hand side takes the form of Newton's difference quotient and can be simplified to the probability density function

$$\begin{aligned} \lim_{\substack{N \rightarrow \infty, k \rightarrow \infty \\ k/N \rightarrow 0}} k \rho_{r,k} &= \frac{F_{\eta}\left(\frac{r}{k}\right) - F_{\eta}\left(\frac{r}{k} - \frac{1}{k}\right)}{\frac{1}{k}} \\ &= \frac{d}{d\eta} F_{\eta}\left(\frac{r}{k}\right) = f_{\eta}\left(\frac{r}{k}\right). \end{aligned} \quad (78)$$

## REFERENCES

- [1] P. J. Moreno, P. P. Ho, and N. Vasconcelos, "A kullback-leibler divergence based kernel for svm classification in multimedia applications," in *Advances in neural information processing systems*, 2003.
- [2] A. B. Hamza and H. Krim, "Image registration and segmentation by maximizing the Jensen-Rényi divergence," in *Energy Minimization Methods in Computer Vision and Pattern Recognition*. Springer, 2003, pp. 147–163.
- [3] K. E. Hild, D. Erdogmus, and J. C. Principe, "Blind source separation using Renyi's mutual information," *Signal Processing Letters, IEEE*, vol. 8, no. 6, pp. 174–176, 2001.
- [4] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh, "Clustering with Bregman divergences," *The Journal of Machine Learning Research*, vol. 6, pp. 1705–1749, 2005.
- [5] S. Ali and S. D. Silvey, "A general class of coefficients of divergence of one distribution from another," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 131–142, 1966.
- [6] X. Nguyen, M. J. Wainwright, and M. I. Jordan, "On surrogate loss functions and f-divergences," *The Annals of Statistics*, pp. 876–904, 2009.
- [7] A. O. Hero, B. Ma, O. Michel, and J. Gorman, "Alpha-divergence for classification, indexing and retrieval," *Communication and Signal Processing Laboratory, Technical Report CSPL-328, U. Mich*, 2001.
- [8] I. Ahmad and P.-E. Lin, "A nonparametric estimation of the entropy for absolutely continuous distributions (corresp.)," *Information Theory, IEEE Transactions on*, vol. 22, no. 3, pp. 372–375, 1976.
- [9] L. Györfi and E. C. Van der Meulen, "Density-free convergence properties of various estimators of entropy," *Computational Statistics & Data Analysis*, vol. 5, no. 4, pp. 425–436, 1987.
- [10] A. J. Izenman, "Review papers: recent developments in nonparametric density estimation," *Journal of the American Statistical Association*, vol. 86, no. 413, pp. 205–224, 1991.
- [11] A. O. Hero III and O. Michel, "Robust entropy estimation strategies based on edge weighted random graphs," in *SPIE's International Symposium on Optical Science, Engineering, and Instrumentation*. International Society for Optics and Photonics, 1998, pp. 250–261.

- [12] V. Berisha, A. Wisler, A. O. Hero III, and A. Spanias, "Empirically estimable classification bounds based on a nonparametric divergence measure," *Signal Processing, IEEE Transactions on*, vol. 64, no. 3, pp. 580–591, 2016.
- [13] J. H. Friedman and L. C. Rafsky, "Multivariate generalizations of the wald-wolfowitz and smirnov two-sample tests," *The Annals of Statistics*, pp. 697–717, 1979.
- [14] K. Fukunaga, *Introduction to statistical pattern recognition*. Academic press, 1990.
- [15] J. A. Costa and A. O. Hero, "Geodesic entropic graphs for dimension and entropy estimation in manifold learning," *Signal Processing, IEEE Transactions on*, vol. 52, no. 8, pp. 2210–2221, 2004.
- [16] M. N. Gorla, N. N. Leonenko, V. V. Mergel, and P. L. Novi Inverardi, "A new class of random vector entropy estimators and its applications in testing statistical hypotheses," *Nonparametric Statistics*, vol. 17, no. 3, pp. 277–297, 2005.
- [17] N. Leonenko, L. Pronzato, V. Savani *et al.*, "A class of rényi information estimators for multidimensional densities," *The Annals of Statistics*, vol. 36, no. 5, pp. 2153–2182, 2008.
- [18] N. Leonenko and L. Pronzato, "Correction of "a class of rényi information estimators for multidimensional densities"," *The Annals of Statistics*, 2010.
- [19] B. Póczos, C. Szepesvári, and D. Tax, "Estimation of rényi entropy and mutual information based on generalized nearest-neighbor graphs," in *Advances in Neural Information Processing Systems*, 2010, pp. 1849–1857.
- [20] B. Póczos, S. Kirshner, and C. Szepesvári, "Rego: Rank-based estimation of rényi information using euclidean graph optimization," in *International Conference on Artificial Intelligence and Statistics*, 2010, pp. 605–612.
- [21] J. Silva and S. S. Narayanan, "Information divergence estimation based on data-dependent partitions," *Journal of Statistical Planning and Inference*, vol. 140, no. 11, pp. 3180–3198, 2010.
- [22] G. A. Darbellay, I. Vajda *et al.*, "Estimation of the information by an adaptive partitioning of the observation space," *IEEE Transactions on Information Theory*, vol. 45, no. 4, pp. 1315–1321, 1999.
- [23] S. Geman and C.-R. Hwang, "Nonparametric maximum likelihood estimation by the method of sieves," *The Annals of Statistics*, pp. 401–414, 1982.
- [24] U. Grenander and G. Ulf, "Abstract inference," Tech. Rep., 1981.
- [25] K. Sricharan, R. Raich, and A. O. Hero, "Estimation of nonlinear functionals of densities with confidence," *Information Theory, IEEE Transactions on*, vol. 58, no. 7, pp. 4135–4159, 2012.
- [26] K. Moon and A. Hero, "Multivariate f-divergence estimation with confidence," in *Advances in Neural Information Processing Systems*, 2014, pp. 2420–2428.
- [27] A. O. Hero III, B. Ma, O. J. Michel, and J. Gorman, "Applications of entropic spanning graphs," *Signal Processing Magazine, IEEE*, vol. 19, no. 5, pp. 85–95, 2002.
- [28] X. Nguyen, M. J. Wainwright, and M. I. Jordan, "Estimating divergence functionals and the likelihood ratio by convex risk minimization," *Information Theory, IEEE Transactions on*, vol. 56, no. 11, pp. 5847–5861, 2010.
- [29] Q. Wang, S. R. Kulkarni, and S. Verdú, "Divergence estimation for multidimensional densities via  $k$ -nearest-neighbor distances," *IEEE Transactions on Information Theory*, vol. 55, no. 5, pp. 2392–2405, 2009.
- [30] —, "Universal estimation of information measures for analog sources," *Foundations and Trends in Communications and Information Theory*, vol. 5, no. 3, pp. 265–353, 2009.
- [31] B. Póczos and J. G. Schneider, "On the estimation of alpha-divergences," in *International Conference on Artificial Intelligence and Statistics*, 2011, pp. 609–617.
- [32] H. Chernoff, "A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations," *The Annals of Mathematical Statistics*, pp. 493–507, 1952.
- [33] A. Bhattacharyya, "On a measure of divergence between two multinomial populations," *Sankhyā: The Indian Journal of Statistics*, pp. 401–406, 1946.
- [34] T. Kailath, "The divergence and bhattacharyya distance measures in signal selection," *Communication Technology, IEEE Transactions on*, vol. 15, no. 1, pp. 52–60, 1967.
- [35] W. A. Hashlamoun, P. K. Varshney, and V. Samarasekera, "A tight upper bound on the bayesian probability of error," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 16, no. 2, pp. 220–224, 1994.
- [36] H. Avi-Itzhak and T. Diep, "Arbitrarily tight upper and lower bounds on the Bayesian probability of error," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 1, pp. 89–91, 1996.
- [37] V. Berisha and A. Hero, "Empirical non-parametric estimation of the fisher information," 2014.
- [38] A. Leblanc, "On estimating distribution functions using bernstein polynomials," *Annals of the Institute of Statistical Mathematics*, vol. 64, no. 5, pp. 919–943, 2012.
- [39] B. C. Turnbull and S. K. Ghosh, "Unimodal density estimation using bernstein polynomials," *Computational Statistics & Data Analysis*, vol. 72, pp. 13–29, 2014.
- [40] S. Ghosal, "Convergence rates for density estimation with bernstein polynomials," *Annals of Statistics*, pp. 1264–1280, 2001.
- [41] G. Igarashi and Y. Kakizawa, "On improving convergence rate of bernstein polynomial density estimator," *Journal of Nonparametric Statistics*, vol. 26, no. 1, pp. 61–84, 2014.
- [42] A. Tenbusch, "Two-dimensional bernstein polynomial density estimators," *Metrika*, vol. 41, no. 1, pp. 233–253, 1994.
- [43] G. J. Babu and Y. P. Chaubey, "Smooth estimation of a distribution and density function on a hypercube using bernstein polynomials for dependent random vectors," *Statistics & probability letters*, vol. 76, no. 9, pp. 959–969, 2006.
- [44] A. Hero and O. J. Michel, "Estimation of rényi information divergence via pruned minimal spanning trees," in *Higher-Order Statistics, 1999. Proceedings of the IEEE Signal Processing Workshop on*. IEEE, 1999, pp. 264–268.
- [45] G. G. Lorentz, *Bernstein polynomials*. American Mathematical Soc., 2012.
- [46] S. N. Bernstein, "Démonstration du théorème de weierstrass fondée sur le calcul des probabilités," *Communications of the Kharkov Mathematical Society*, vol. 13, pp. 1–2, 1912.
- [47] J. A. Gubner, *Probability and random processes for electrical and computer engineers*. Cambridge University Press, 2006.
- [48] D. J. Sutherland, L. Xiong, B. Póczos, and J. Schneider, "Kernels on sample sets via nonparametric divergence estimates," *arXiv preprint arXiv:1202.0302*, 2012.
- [49] Z. Szabó, "Information theoretical estimators toolbox," *Journal of Machine Learning Research*, vol. 15, pp. 283–287, 2014.
- [50] S. Kullback, *Information theory and statistics*. Courier Corporation, 1997.
- [51] A. Wisler, V. Berisha, K. Ramamurthy, D. Wei, and A. Spanias, "Empirically-estimable multi-class performance bounds," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2016.
- [52] K. Fukunaga, *Introduction to statistical pattern recognition*. Academic press, 1990.
- [53] L. Devroye, L. Györfi, and G. Lugosi, *A probabilistic theory of pattern recognition*. Springer Science & Business Media, 2013, vol. 31.